

Tighter Lower Bounds on Mutual Information for Fiber-Optic Channels

Naga V. Irukulapati, *Student Member, IEEE*, Marco Secondini, *Member, IEEE*, Erik Agrell, *Senior Member, IEEE*, Pontus Johannisson, and Henk Wymeersch, *Member, IEEE*

Abstract—In fiber-optic communications, evaluation of mutual information (MI) is still an open issue due to the unavailability of an exact and mathematically tractable channel model. Traditionally, lower bounds on MI are computed by approximating the (original) channel with an auxiliary forward channel. In this paper, lower bounds are computed using an auxiliary backward channel, which has not been previously considered in the context of fiber-optic communications. Distributions obtained through two variations of the stochastic digital backpropagation (SDBP) algorithm are used as auxiliary backward channels and these bounds are compared with bounds obtained through the conventional digital backpropagation (DBP). Through simulations, higher information rates were achieved with SDBP compared with DBP, which implies that tighter lower bound on MI can be achieved through SDBP.

Index Terms—Achievable information rate, auxiliary channel, fiber-optical communications, mismatched decoding, nonlinear compensation, stochastic digital backpropagation.

I. INTRODUCTION

Shannon proved that reliable communication through a noisy channel is possible with channel coding, as long as the information rate is less than the channel capacity [1]. Reliable communication means that coding schemes exist that can make the probability of error arbitrarily small. Furthermore, if the information rate is greater than the channel capacity, then regardless of the coding scheme, the probability of error cannot be made zero. The channel capacity for an additive white Gaussian noise (AWGN) channel is known exactly and has been derived in [1, Sec. 24]. However, the evaluation of the channel capacity for the fiber-optic channel (FOC) accounting for the dispersion, nonlinearity, and noise is still an open problem due to the unavailability of an exact and mathematically tractable channel model. Therefore, accurately predicting the capacity of the FOC has been the focus of much recent research [2]–[7].

For a fixed input distribution, the mutual information (MI) gives a lower bound on the channel capacity. MI is also used for predicting the post-forward-error-correction (FEC) bit-error rate (BER) between the input and output of the discrete-

time channel. MI is shown to be a better metric than the pre-FEC BER for estimating the post-FEC BER in soft-decision FEC systems [8]–[11]. For the FOC, the most commonly used approach to lower-bound the MI is to approximate the original forward channel with an auxiliary forward channel (AFC) [11]–[17]. A receiver that is optimal for an AFC is used to process the data generated from the original forward channel and to compute an information rate. This rate is achievable by that receiver and, for this reason, is often referred to as an achievable information rate (AIR) [14], [18], [19]. However, in some scenarios, only an auxiliary backward channel (ABC)—i.e., an approximation for the posterior distribution of the channel input given its output—is available, while a corresponding AFC does not exist or is computationally too complex to find. In such cases, an alternative approach to lower-bound the MI is through a direct use of the ABC. This approach will be applied in the context of the FOC for the first time in this paper. The concept of ABC was used for the first time (to the best of our knowledge) in the context of universal decoding for memoryless channels with deterministic interference [20]. An ABC is used instead of an AFC to maximize the lower bound using an iterative procedure [21]. A receiver that is optimal for an ABC is used to process the data generated from the original forward channel and to compute an AIR that is achievable by that receiver. In this paper, distributions obtained from two variations of the stochastic digital backpropagation (SDBP) algorithm are used as ABCs, namely from symbol-by-symbol SDBP (SBS-SDBP) [22] and Gaussian message passing SDBP (GMP-SDBP) [23]. Through simulations, the AIR computed using these two distributions of SDBP was observed to be higher than the AIR obtained using the conventional digital backpropagation (DBP) algorithm, implying that a tighter lower bound on the MI can be obtained using SDBP.

Organization of the paper: MI is mathematically introduced in Sec. II. Lower bounds on the MI using an AFC and an ABC and their Monte Carlo (MC) estimation are described in Sec. III. Computation of AIRs for the FOC is considered in Sec. IV, where SBS-SDBP and GMP-SDBP are described briefly, leading to a discussion on how ABCs are obtained using these two approaches. AIRs computed using these two versions of SDBP are then compared with DBP. Numerical results are presented and discussed in Sec. V, followed by conclusions in Sec. VI.

II. MUTUAL INFORMATION

For a discrete-time channel with memory, the MI between random vectors $\mathbf{X} \triangleq (X_1, X_2, \dots, X_K)$ and $\mathbf{Y} \triangleq$

N. V. Irukulapati, H. Wymeersch, and E. Agrell are with the Dept. of Signals and Systems at Chalmers University of Technology, 41296 Gothenburg Sweden. M. Secondini is with the Institute of Communication, Information, and Perception Technologies, Scuola Superiore Sant'Anna, Pisa, Italy. P. Johannisson is with Acreo Swedish ICT AB, 40014 Gothenburg, Sweden. Corresponding author's email: vnaga@chalmers.se

This research was supported by the Swedish Research Council (VR) under grant 2013-5642 and the European Research Council under grant no. 258418 (COOPNET). The simulations were performed in part on resources provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE.

(Y_1, Y_2, \dots, Y_J) with $J \geq K$ is defined as¹

$$I(\mathbf{X}; \mathbf{Y}) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} d\mathbf{x} d\mathbf{y} \\ \triangleq \mathbb{E}_{\mathbf{X}, \mathbf{Y}} \left[\log \frac{p(\mathbf{Y}|\mathbf{X})}{p(\mathbf{Y})} \right], \quad (1)$$

where $\mathbf{x} = (x_1, x_2, \dots, x_K) \in \mathcal{X}$ is a realization of \mathbf{X} drawn from the input distribution $p(\mathbf{x})$, and $\mathbf{y} = (y_1, y_2, \dots, y_J) \in \mathcal{Y}$ is a realization of the corresponding output random vector \mathbf{Y} . $p(\mathbf{y}|\mathbf{x})$ is the channel conditional distribution, $p(\mathbf{y}) = \int_{\mathcal{X}} p(\mathbf{x})p(\mathbf{y}|\mathbf{x})d\mathbf{x}$ is the output distribution, and $\mathbb{E}_{\mathbf{X}, \mathbf{Y}}\{\cdot\}$ is expectation over the joint distribution $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$.

Alternatively, the MI can also be written as

$$I(\mathbf{X}; \mathbf{Y}) = \mathbb{E}_{\mathbf{X}, \mathbf{Y}} \left[\log \frac{p(\mathbf{X}|\mathbf{Y})}{p(\mathbf{X})} \right], \quad (2)$$

where $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})/p(\mathbf{y})$. The information rate between the ergodic processes for the channel with memory is [19]

$$I^{\text{mem}} = \lim_{K, J \rightarrow \infty} \frac{1}{K} I(\mathbf{X}; \mathbf{Y}). \quad (3)$$

It is often the case, especially for the FOC, that the channel distributions, $p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{y})$, are not known in closed form. Hence, the MI of (1) and (2) and subsequently the information rate of (3) cannot be computed in closed form. The information rate of (3) can, in principle, be estimated through simulations using the BCJR algorithm [19]. However, the complexity of this simulation-based technique increases exponentially with memory and, hence, is infeasible for systems with a large memory such as the FOC [24]. In such instances, upper and lower bounds on (3) are calculated. A useful and practical approach to lower-bound the MI is by using a concept known as mismatched decoding [25], [26]. In mismatched decoding, the original distributions $p(\mathbf{y}|\mathbf{x})$ or $p(\mathbf{x}|\mathbf{y})$ are approximated with auxiliary distributions and the rates computed using these auxiliary distributions are a lower bound on the MI, as will be detailed in Sec. III-A and III-B. The better the auxiliary distribution approximates the original distribution, the tighter are these bounds. Currently, all lower bounds on the MI for the FOC are based on AFCs [14]–[16]. In this paper, we use an ABC to compute lower bounds on the MI for the FOC and also investigate if the existing bounds can be further improved, thereby obtaining tighter bounds on the MI.

Before proceeding, we define four entities, similar to [21, Fig. 1], that will be used throughout the paper:

- $p(\mathbf{y}|\mathbf{x})$ is the original forward channel;
- $p(\mathbf{x}|\mathbf{y})$ is the original backward channel;
- $q(\mathbf{y}|\mathbf{x})$ is an AFC;
- $r(\mathbf{x}|\mathbf{y})$ is an ABC.

We define a backward channel by reversing the usual meaning

of \mathbf{X} and \mathbf{Y} , i.e., looking at \mathbf{X} as being the output of some channel which is fed by \mathbf{Y} , which in turn is produced by some source [20], [21]. The input and output alphabets of the AFC match the input and output alphabets of the original forward channel. Similarly, the input and output alphabets of the ABC match the input and output alphabets of the original backward channel. Note that original backward channel is associated with the original forward channel as $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})/p(\mathbf{y})$. However, it is not necessary that such a relation exists between auxiliary channels, i.e., $r(\mathbf{x}|\mathbf{y})$ can be any conditional distribution which does not correspond² to any AFC $q(\mathbf{y}|\mathbf{x})$.

III. LOWER BOUNDS ON MUTUAL INFORMATION

Lower bounds using an AFC $q(\mathbf{y}|\mathbf{x})$ and an ABC $r(\mathbf{x}|\mathbf{y})$ have been derived in [18], [19], [21] and are briefly described here for completeness.

A. Lower Bounds using Auxiliary Forward Channel $q(\mathbf{y}|\mathbf{x})$

For any input distribution $p(\mathbf{x})$, original forward channel $p(\mathbf{y}|\mathbf{x})$, and AFC $q(\mathbf{y}|\mathbf{x})$, a lower bound on the MI is [19, Eq. (41)]

$$I(\mathbf{X}; \mathbf{Y}) \geq I_q(\mathbf{X}; \mathbf{Y}) \triangleq \mathbb{E}_{\mathbf{X}, \mathbf{Y}} \left[\log \frac{q(\mathbf{Y}|\mathbf{X})}{q(\mathbf{Y})} \right] \\ = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(\mathbf{x}, \mathbf{y}) \log \frac{q(\mathbf{y}|\mathbf{x})}{q(\mathbf{y})} d\mathbf{x} d\mathbf{y}, \quad (4)$$

where $q(\mathbf{y}) \triangleq \int_{\mathcal{X}} p(\mathbf{x})q(\mathbf{y}|\mathbf{x})d\mathbf{x}$ is the output distribution obtained by connecting the original source $p(\mathbf{x})$ to the AFC. It has been shown that the lower bound, $I_q(\mathbf{X}; \mathbf{Y})$ of (4), can be achieved by using an optimal detector, i.e., a maximum a posteriori detector, designed for the AFC and used as a receiver for the original forward channel [18], [25], i.e., the decisions $\hat{\mathbf{x}}$ are taken as

$$\hat{\mathbf{x}}(\mathbf{y}) = \arg \max_{\mathbf{x}} q(\mathbf{y}|\mathbf{x})p(\mathbf{x}). \quad (5)$$

By achievability, we mean that if the data rate is lower than the rate calculated using (4), there exist coding schemes that can make the error probability arbitrarily low. Since the data generated by the original forward channel is processed by a receiver that is optimized for a different AFC, this approach is known as mismatched decoding.

Using the Kullback–Leibler divergence (KLD) [27, Sec. 8.1], it can be easily verified [19, Eq. (34)–(41)] that (4) is indeed a lower bound,

$$I(\mathbf{X}; \mathbf{Y}) - I_q(\mathbf{X}; \mathbf{Y}) = \mathbb{E}_{\mathbf{X}, \mathbf{Y}} \left[\log \frac{p(\mathbf{Y}|\mathbf{X})}{p(\mathbf{Y})} \frac{q(\mathbf{Y})}{q(\mathbf{Y}|\mathbf{X})} \right] \quad (6)$$

$$= \mathbb{E}_{\mathbf{X}, \mathbf{Y}} \left[\log \frac{p(\mathbf{X}, \mathbf{Y})}{p(\mathbf{Y})r_q(\mathbf{X}|\mathbf{Y})} \right] \quad (7)$$

$$= D(p(\mathbf{x}, \mathbf{y}) || p(\mathbf{y})r_q(\mathbf{x}|\mathbf{y})) \geq 0, \quad (8)$$

where

$$r_q(\mathbf{x}|\mathbf{y}) \triangleq \frac{p(\mathbf{x})q(\mathbf{y}|\mathbf{x})}{q(\mathbf{y})} \quad (9)$$

²To highlight this difference, we chose to use $r(\mathbf{x}|\mathbf{y})$ for an ABC instead of $q(\mathbf{x}|\mathbf{y})$.

¹All logarithms in this paper are in base 2; therefore, MI will be measured in bits. To simplify the notation, we used $p(\mathbf{y}|\mathbf{x})$, $p(\mathbf{x})$, and $p(\mathbf{y})$ instead of explicitly writing $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$, $p_{\mathbf{X}}(\mathbf{x})$, and $p_{\mathbf{Y}}(\mathbf{y})$. So, $p(\mathbf{y}|\mathbf{x})$, $p(\mathbf{x})$, $p(\mathbf{y})$, and their auxiliary counterparts refer to different distributions. For the limit in (3) to exist, we assume the existence of sequences of distributions $p(\mathbf{y}|\mathbf{x})$, $p(\mathbf{x})$, and $p(\mathbf{y})$ for $K = 1, 2, \dots$

is the ABC induced by the AFC $q(\mathbf{y}|\mathbf{x})$. A sufficient condition for the inequality (8) to hold is that $p(\mathbf{y})r_q(\mathbf{x}|\mathbf{y})$ is a joint distribution, i.e., $\int_{\mathcal{X}} \int_{\mathcal{Y}} p(\mathbf{y})r_q(\mathbf{x}|\mathbf{y})d\mathbf{x}d\mathbf{y} = 1$ [27, Th. 8.6.1]. This condition is fulfilled for any combination of $p(\mathbf{y}|\mathbf{x})$ and $q(\mathbf{y}|\mathbf{x})$, and can be verified by using $q(\mathbf{y}) = \int_{\mathcal{X}} p(\mathbf{x})q(\mathbf{y}|\mathbf{x})d\mathbf{x}$ in $r_q(\mathbf{x}|\mathbf{y})$ of (9).

B. Lower Bounds using Auxiliary Backward Channel $r(\mathbf{x}|\mathbf{y})$

There are instances such as SDBP, where an ABC $r(\mathbf{x}|\mathbf{y})$ is known while the corresponding AFC $q(\mathbf{y}|\mathbf{x})$ is unknown. In such cases, if (4) is to be used to compute a lower bound on the MI, an AFC $q(\mathbf{y}|\mathbf{x})$ has to be computed corresponding to a given ABC $r(\mathbf{x}|\mathbf{y})$ and a given $p(\mathbf{x})$. There are two challenges with this approach. Firstly, given $r(\mathbf{x}|\mathbf{y})$ and $p(\mathbf{x})$, no general method exists to compute $q(\mathbf{y}|\mathbf{x})$, and depending on the input and output alphabets, there may not always exist a corresponding $q(\mathbf{y}|\mathbf{x})$ or there may exist multiple solutions. Secondly, even if a solution exists to get an AFC corresponding to an ABC and $p(\mathbf{x})$, it may be computationally very difficult to compute it. In this section, we provide an alternate approach to lower-bound the MI using $r(\mathbf{x}|\mathbf{y})$, without the explicit knowledge of a corresponding AFC.

For any input distribution $p(\mathbf{x})$ and any conditional distribution $r(\mathbf{x}|\mathbf{y})$, lower bounds on the MI can be derived as [21, Eq. (39)]

$$I(\mathbf{X}; \mathbf{Y}) \geq I_r(\mathbf{X}; \mathbf{Y}) \triangleq \mathbb{E}_{\mathbf{X}, \mathbf{Y}} \left[\log \frac{r(\mathbf{X}|\mathbf{Y})}{p(\mathbf{X})} \right], \quad (10)$$

where, similarly to (4), the averaging is computed with respect to the joint distribution $p(\mathbf{x}, \mathbf{y})$. The lower bound (10) can be also proved using KLD as

$$I(\mathbf{X}; \mathbf{Y}) - I_r(\mathbf{X}; \mathbf{Y}) = \mathbb{E}_{\mathbf{X}, \mathbf{Y}} \left[\log \frac{p(\mathbf{X}, \mathbf{Y})}{p(\mathbf{Y})r(\mathbf{X}|\mathbf{Y})} \right] \quad (11)$$

$$= D(p(\mathbf{x}, \mathbf{y}) || p(\mathbf{y})r(\mathbf{x}|\mathbf{y})) \geq 0 \quad (12)$$

For the inequality in (12) to hold, $p(\mathbf{y})r(\mathbf{x}|\mathbf{y})$ should be a joint distribution, which is always true as long as $r(\mathbf{x}|\mathbf{y})$ is chosen as a conditional distribution. Note that if the ABC is given by $r_q(\mathbf{x}|\mathbf{y})$, i.e., induced by the AFC according to (9), then (4) and (10) provide the same lower bound, i.e., $I_q(\mathbf{X}; \mathbf{Y}) = I_r(\mathbf{X}; \mathbf{Y})$. However, it should be noted that (10) can be used as a lower bound on the MI for any arbitrary conditional distribution $r(\mathbf{x}|\mathbf{y})$, which is not necessarily related to any $q(\mathbf{y}|\mathbf{x})$. It has been shown that $I_r(\mathbf{X}; \mathbf{Y})$ is achievable using an optimal detector, i.e., a maximum a posteriori detector, designed for the ABC $r(\mathbf{x}|\mathbf{y})$ [18], [21, Eq. (42)–(43)]. The decisions $\hat{\mathbf{x}}$ are taken as

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x}} r(\mathbf{x}|\mathbf{y}). \quad (13)$$

Remark 1: Either maximizing I_q of (4) over all possible AFCs $q(\mathbf{y}|\mathbf{x})$ or maximizing I_r of (10) over all possible ABCs $r(\mathbf{x}|\mathbf{y})$ leads to the true MI (1)–(2).

C. Monte Carlo Estimation of AIR

By combining (3) and (4), and (3) and (10), we have

$$I_q^{\text{mem}} \triangleq \lim_{K, J \rightarrow \infty} \frac{1}{K} I_q(\mathbf{X}; \mathbf{Y}) = \lim_{K, J \rightarrow \infty} \frac{1}{K} \mathbb{E}_{\mathbf{X}, \mathbf{Y}} \left[\log \frac{q(\mathbf{Y}|\mathbf{X})}{q(\mathbf{Y})} \right], \quad (14)$$

$$I_r^{\text{mem}} \triangleq \lim_{K, J \rightarrow \infty} \frac{1}{K} I_r(\mathbf{X}; \mathbf{Y}) = \lim_{K, J \rightarrow \infty} \frac{1}{K} \mathbb{E}_{\mathbf{X}, \mathbf{Y}} \left[\log \frac{r(\mathbf{X}|\mathbf{Y})}{p(\mathbf{X})} \right]. \quad (15)$$

The state-of-the-art method for the estimation of (14) and (15) is a simulation based on MC averages [19], [28]. This method utilizes the asymptotic equipartition property for ergodic processes, which states that, for long enough sequence length K , $\log(p(\mathbf{a}))/K$ for a single realization \mathbf{a} converges to the expectation $E_{\mathbf{A}}[\log(p(\mathbf{a}))/K]$ with probability one as $K \rightarrow \infty$ [27, Ch. 3]. The channel is simulated N_{mc} times, each time by generating input \mathbf{x} with different random seed, say $\mathbf{x}^{(n)}$ for the n th MC run, to get a corresponding output $\mathbf{y}^{(n)}$ from the FOC for each MC run. The lower bound on the MIs (14) and (15) can then be estimated as

$$\hat{I}_q^{\text{mem}} = \frac{1}{N_{\text{mc}}} \sum_{n=1}^{N_{\text{mc}}} \left\{ \frac{1}{K} \log \frac{q(\mathbf{y}^{(n)}|\mathbf{x}^{(n)})}{q(\mathbf{y}^{(n)})} \right\}, \quad (16)$$

$$\hat{I}_r^{\text{mem}} = \frac{1}{N_{\text{mc}}} \sum_{n=1}^{N_{\text{mc}}} \left\{ \frac{1}{K} \log \frac{r(\mathbf{x}^{(n)}|\mathbf{y}^{(n)})}{p(\mathbf{x}^{(n)})} \right\}. \quad (17)$$

IV. COMPUTATION OF AIR FOR THE FOC

The computation of AIRs for the FOC using different auxiliary channels is abstracted in Fig. 1. The input data \mathbf{x} is sent through a pulse shaper followed by the fiber link to get the output of the FOC, \mathbf{y} . This output is fed either to SDBP or DBP, which aim to undo the impairments induced by the FOC. The effect of pulse shaping is reversed using either the output from DBP or through one of the two techniques for SDBP to get the auxiliary channels. In the section, these processes will be detailed.

A. Computation of AIR using DBP

The traditional approach of computing the AIR is by assuming the AFC to be memoryless³. This assumption is justified by using a post-processing block, such as a DBP block for nonlinear compensation, after the FOC as part of the AFC. The output statistics of the FOC and the DBP has been considered memoryless with additive Gaussian noise with the same variance in all dimensions [14], [15]. Recently, the correlation between the in-phase and quadrature components was accounted for in the Gaussian assumption [16]. In that study, it was shown that significant gains in the AIR are possible depending on the scenarios such as inline dispersion compensation at high powers. We will use both these approaches in benchmarking the results. As shown in Fig. 1, let $\mathbf{z} = [z_1, z_2, \dots, z_K]$, with $z_i \in \mathbb{C}$ for $i = 1, 2, \dots, K$, be the signal after DBP, matched filtering (MF) and sampling.

³When the output of the channel y_i at discrete time i , given the channel input x_i at time i , is independent of channel inputs and outputs at all other times, we call the channel as memoryless.

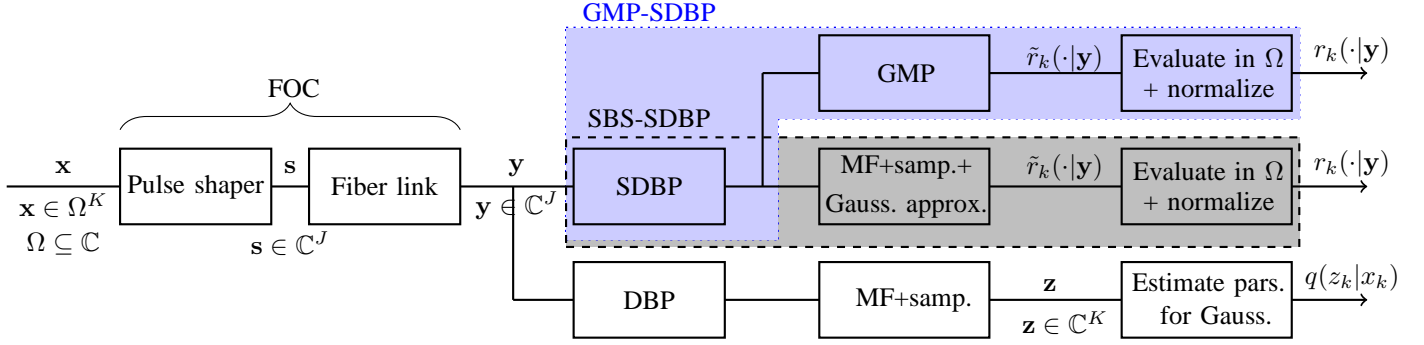


Fig. 1. Auxiliary channels obtained for the FOC using two variations of SDBP, and DBP, where $x_k \in \Omega$.

According to the data-processing inequality [27, Ch. 2], the information content of a signal cannot be increased after post-processing and hence we have $I(\mathbf{X}; \mathbf{Z}) \leq I(\mathbf{X}; \mathbf{Y})$, where

$$I(\mathbf{X}; \mathbf{Z}) = \mathbb{E}_{\mathbf{X}, \mathbf{Z}} \left[\log \frac{p(\mathbf{Z}|\mathbf{X})}{p(\mathbf{Z})} \right]$$

with $p(\mathbf{z}) = \int_{\mathbf{x}} p(\mathbf{x})p(\mathbf{z}|\mathbf{x})d\mathbf{x}$. Similar to the lower bound (4) using AFC, we can define

$$I_q(\mathbf{X}; \mathbf{Z}) \triangleq \mathbb{E}_{\mathbf{X}, \mathbf{Z}} \left[\log \frac{q(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right] \leq I(\mathbf{X}; \mathbf{Z}), \quad (18)$$

with $q(\mathbf{z}) = \int_{\mathbf{x}} p(\mathbf{x})q(\mathbf{z}|\mathbf{x})d\mathbf{x}$. Using DBP as detector, $q(\mathbf{z}|\mathbf{x})$ is commonly assumed to factorize into marginal distributions, i.e., $q(\mathbf{z}|\mathbf{x}) = \prod_{k=1}^K q(z_k|x_k)$ and $q(\mathbf{z}) = \prod_{k=1}^K q(z_k)$. By using these factorizations in (18) and using an equivalent of (16) for \mathbf{z} as output, we have

$$\hat{I}_q(X; Z) = \frac{1}{N_{mc}} \sum_{n=1}^{N_{mc}} \left\{ \frac{1}{K} \sum_{k=1}^K \log \frac{q(z_k^{(n)}|x_k^{(n)})}{q(z_k^{(n)})} \right\}, \quad (19)$$

where, similarly to the approach in [16], $q(z_k|x_k)$ is assumed⁴ to be a Gaussian distribution with a different mean and covariance matrix for each possible value of x_k . In particular, the real and imaginary components of z_k are taken to be either independent and identically distributed Gaussian (iidG) or correlated Gaussian (CG). In both these variations, a training phase is employed to obtain mean and variance corresponding to each of the constellation points. This training phase is visualized as the ‘Estimate pars. for Gauss.’ block in Fig. 1, referring to the estimation of parameters for the Gaussian distribution. The means are obtained using [16, Eq. (8)], and variances for iidG and CG are obtained using [16, Eq. (9)] and [16, Eq. (10)], respectively.

Remark 2: Note that even though AIRs for DBP are computed using an AFC, the same AIRs are also obtained using an ABC induced by this AFC, i.e., $q(\mathbf{z}|\mathbf{x})p(\mathbf{x})/q(\mathbf{z})$.

B. Computation of AIR using SDBP

DBP is not an optimal processing strategy for nonlinear compensation. Indeed, some residual memory due to signal–

noise interaction is present even after DBP is performed. As SDBP accounts for this memory, it may lead to tighter bounds on the MI.

The theory behind SDBP is based on factor graphs and message passing, and is derived and explained in detail in [22], while improved versions of SDBP are found in [23], [29], [30]. A short summary of SDBP is provided here for completeness. SDBP compensates not only for linear and nonlinear effects existing in the fiber but also accounts for the noise from the amplifiers. The main idea of SDBP is to represent the uncertainty present in the unobserved signals at each stage of the FOC. These unobserved signals are signals after each of the linear and nonlinear blocks of the split-step Fourier method (SSFM) and also the signals after the amplifiers. Uncertainty is captured through a collection of N_p waveforms, which represents a distribution. These N_p waveforms are passed through the inverse of each of the blocks of the FOC, starting from the received signal \mathbf{y} . SDBP can thus be viewed as an algorithm that takes an input $\mathbf{y} \in \mathbb{C}^J$ and returns a vector of N_p outputs, where each of these N_p outputs is in \mathbb{C}^J , and describes the knowledge the receiver has regarding the variable \mathbf{s} in Fig. 1. To account for the effect of pulse shaping, two different approaches, SBS-SDBP and GMP-SDBP, are proposed in [22] and [23], respectively, and are explained briefly below.

In the first approach, SBS-SDBP, the output after SDBP is passed through MF followed by sampling⁵. Corresponding to each symbol x_k , N_p outputs are approximated with a multivariate Gaussian distribution, $\tilde{r}_k(\cdot|\mathbf{y})$. In the second approach, referred to as GMP-SDBP in this paper, all N_p outputs from SDBP are first approximated with a multivariate Gaussian distribution and then Gaussian message passing (GMP) is applied according to [33, Table III] instead of an MF, and $\tilde{r}_k(\cdot|\mathbf{y})$ is obtained. In SBS-SDBP and GMP-SDBP, the distribution $\tilde{r}_k(\cdot|\mathbf{y})$ is evaluated at $x_k \in \Omega$ and normalized to get an ABC as shown in Fig. 1. By assuming that the input distribution $p(\mathbf{x})$ is the product of its marginals, i.e., $p_{\mathbf{X}}(\mathbf{x}) = \prod_{k=1}^K p_X(x_k)$ and by assuming that the ABC is

⁵There is residual memory left after SBS-SDBP [29] as MF followed by sampling is a linear technique and may not be the optimal processing for the nonlinear FOC [31], [32]. This residual memory was accounted for by using the Viterbi algorithm on the samples obtained after MF, and was shown to have improved performance compared to SBS-SDBP [30].

⁴When superscripts for indicating MC run n are omitted as in $q(z_k|x_k)$, it should be interpreted as applying to any general MC run.

factorized as

$$r(\mathbf{x}|\mathbf{y}) \triangleq \prod_{k=1}^K r_k(x_k|\mathbf{y}), \quad (20)$$

(17) becomes

$$\hat{I}_r^{\text{mem}} = \frac{1}{N_{\text{mc}}} \sum_{n=1}^{N_{\text{mc}}} \left\{ \frac{1}{K} \sum_{k=1}^K \log \frac{r_k(x_k^{(n)}|\mathbf{y}^{(n)})}{p_X(x_k^{(n)})} \right\}, \quad (21)$$

where $r_k(x_k^{(n)}|\mathbf{y}^{(n)})$ is obtained by either SBS-SDBP or GMP-SDBP.

Effect of Number of Particles in SDBP: Given an output sequence \mathbf{y} , SDBP provides an ABC $r_k(\cdot|\mathbf{y})$. However, in practice, this distribution is dependent on the number of particles, N_p . Specifically, the statistics of $r_k(\cdot|\mathbf{y})$ are better captured with a high number of particles, and the statistics do not change significantly after a certain $N_p > N'_p$. In other words, when $N_p < N'_p$, distributions obtained from SDBP can be dependent on the particular noise realization used in SDBP. However, the bounds obtained are still valid although they are not as tight as with $N_p > N'_p$. In order to prove this, we consider the noise realizations added in SDBP $\mathbf{w} \in \mathbb{C}^{J \times N_p \times N}$ as an additional variable independent of \mathbf{x} provided by the detector itself, where N is the number of spans in the fiber link. The ABC provided by SDBP can then be written as $r_k(\cdot|\mathbf{y}, \mathbf{w})$. By plugging this in the ABC lower bound (10), an achievable lower bound $I_r(\mathbf{X}; \mathbf{Y}, \mathbf{W})$ can be computed between the input \mathbf{X} and output pairs \mathbf{Y}, \mathbf{W} as

$$I_r(\mathbf{X}; \mathbf{Y}, \mathbf{W}) = \iiint p(\mathbf{x}, \mathbf{y}, \mathbf{w}) \log \frac{r(\mathbf{x}|\mathbf{y}, \mathbf{w})}{p(\mathbf{x})} d\mathbf{x} d\mathbf{y} d\mathbf{w}, \quad (22)$$

$$= \int p(\mathbf{w}) d\mathbf{w} \iint p(\mathbf{x}, \mathbf{y}) \log \frac{r(\mathbf{x}|\mathbf{y}, \mathbf{w})}{p(\mathbf{x})} d\mathbf{x} d\mathbf{y}, \quad (23)$$

$$= \int p(\mathbf{w}) d\mathbf{w} I_{r_{\mathbf{w}}}(\mathbf{X}; \mathbf{Y}) \quad (24)$$

$$\leq \int p(\mathbf{w}) d\mathbf{w} I(\mathbf{X}; \mathbf{Y}) = I(\mathbf{X}; \mathbf{Y}), \quad (25)$$

where in (23), we factorized the joint distribution due to the fact that \mathbf{w} does not provide any additional information about \mathbf{x} (as it is independent of \mathbf{x} given \mathbf{y}). $r_{\mathbf{w}}$ in (24) is to explicitly show that the ABC is dependent on \mathbf{w} and the inequality in (25) follows from (10) for any ABC r . It can be noticed that the inequality in (25) is valid irrespective of N_p .

V. NUMERICAL RESULTS AND DISCUSSION

A. System Parameters

The FOC used in this paper is identical to [30, Fig. 1] with a single-channel system comprising a single-polarization transmitter⁶ and a fiber link consisting of N spans. Each span of the fiber link consists of a transmission fiber of length L , which is a standard single-mode fiber (SMF) simulated using SSFM, and a fiber Bragg grating (FBG) for optical

⁶The theory remains the same for dual-polarization FOC also. As GMP-SDBP involves inverting a very high-dimensional matrix, we simulated only for single-polarization in this paper.

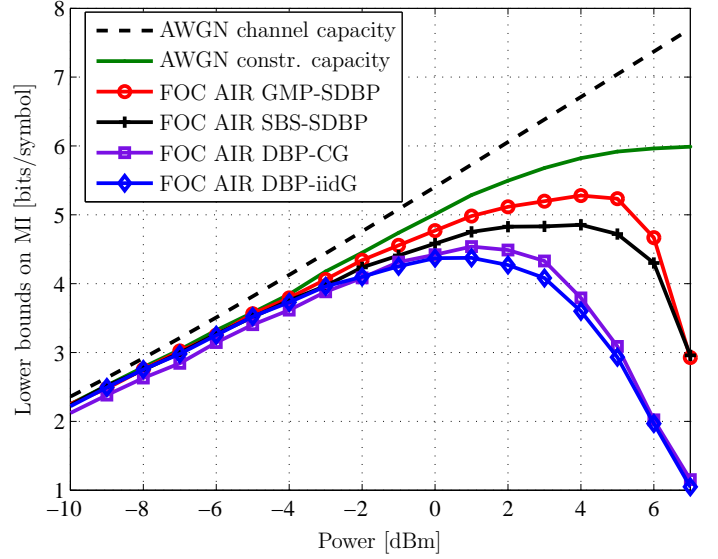


Fig. 2. Lower bounds on the MI using DBP-iidG, DBP-CG, SBS-SDBP, and GMP-SDBP for 14 GBd, 64-QAM, FBG link, $N = 30$, $L = 120$ km.

dispersion-managed (DM) links. In between fiber spans, there are EDFAs that compensate for the losses in the fiber. The transmitter uses a root raised cosine pulse shaper with a roll-off factor of 0.25 and truncation length of 16 symbol periods. The modulation format is 64-QAM and the symbol rate, R_s , is either 14 GBd or 28 GBd. The parameters used for the SMF are a dispersion coefficient of $D = 16$ ps/(nm km), a Kerr nonlinearity parameter of $\gamma = 1.3$ (W km)⁻¹, and an attenuation of 0.2 dB/km, which are according to the ITU-T G.652 standard. Propagation in the fiber is simulated using the SSFM with a segment length [34] of $\Delta = (\epsilon L_N L_D^2)^{1/3}$, where $\epsilon = 10^{-4}$, $L_N = 1/(\gamma P)$ is the nonlinear length, $L_D = 2\pi c/(R_s^2 |D| \lambda^2)$ is the dispersion length, λ is the wavelength, c is the speed of light, and P is the average input power to each fiber span. We used the same segment lengths for simulating the channel and for both DBP and SDBP. An FBG with an insertion loss of 3 dB and perfect dispersion compensation for the preceding SMF is used. The noise figure is 5.5 dB for each of the amplifiers. Ideal band-pass filters with an equivalent low-pass bandwidth equal to the symbol rate are used in the EDFAs and at the input of the receiver. The number of particles used in the SDBP approach is $N_p = 500$ for both SBS-SDBP and GMP-SDBP. The input distribution is assumed to be uniform, i.e., $p_X(x_k) = 1/|\Omega|$ for $x_k \in \Omega$.

B. Results

Fig. 2 shows the lower bounds on the MI, AIR, as a function of input power, obtained through different auxiliary channels for 14 GBd over a 64-QAM FBG link. For reference, the capacity of the AWGN channel and the constrained capacity of an equivalent AWGN channel with 64-QAM are also shown. We notice a trend for the DBP approaches (diamond and square markers) to have lower AIRs than the SDBP approaches (plus and circle markers), and also that the AIRs behave differently for DBP and SDBP. Specifically, we can observe from the DBP curves that up to about 0 dBm power, the

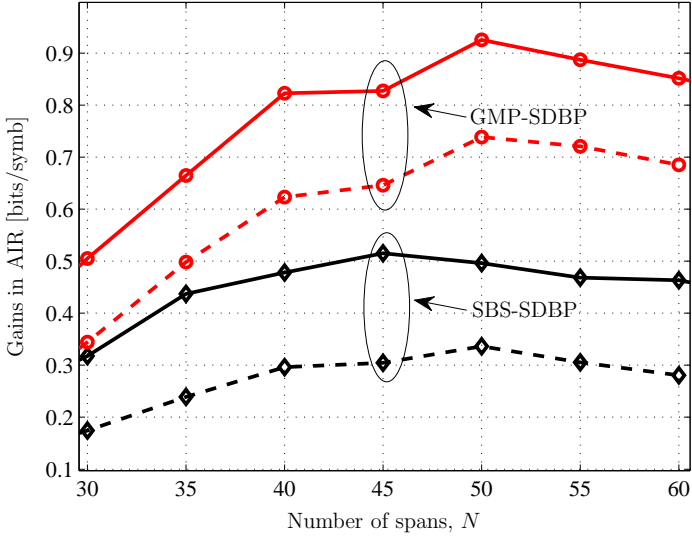


Fig. 3. Gains in AIR for 14 GBd, 64-QAM, $L = 100$ km over DBP-iidG (solid line) and DBP-CG (dashed line).

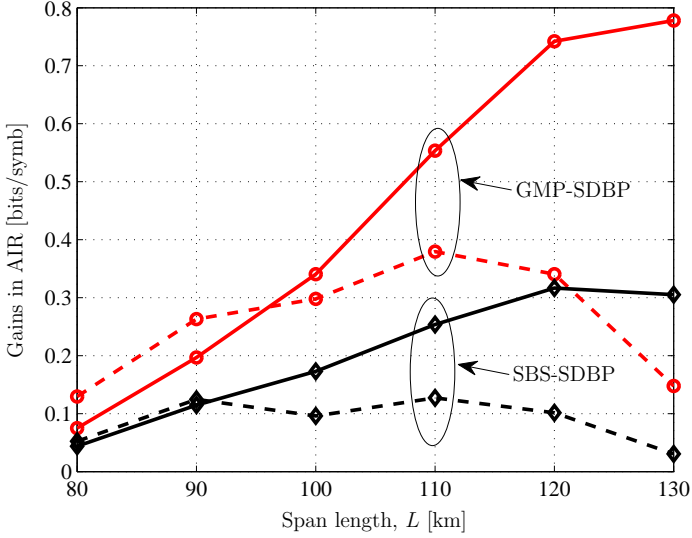


Fig. 4. Gains in AIR for 64-QAM, $N = 30$ over DBP-CG for 14 GBd (solid line) and 28 GBd (dashed line).

behavior of this link is approximately linear. At higher powers (beyond 1 dBm for DBP and 4 dBm for SDBP), nonlinearity comes into play and the AIR decreases. However, the AIR for DBP decreases faster than that of SDBP, which means that SDBP performs better in the nonlinear regime. This is an expected behavior because SDBP accounts for the nonlinear signal-noise interactions, which DBP does not account for. If we compare AIRs between DBP techniques, DBP-CG (square markers) has better AIR than DBP-iidG (diamond markers), as the former accounts for the correlation between the in-phase and quadrature components. Also, we can observe that GMP-SDBP has better AIR than SBS-SDBP, as GMP-SDBP is a more principled way of computing a specific message, whereas SBS-SDBP is a heuristic approach.

Lower bounds on the MI similar to Fig. 2 were computed for

SBS-SDBP, GMP-SDBP, DBP-CG, and DBP-iidG by varying the number of spans for 14 GBd, 64-QAM, and $L = 100$ km. For each N , the maximum AIR is computed for these four approaches, and the gains are plotted in Fig. 3. The DBP-CG and DBP-iidG approaches are used to benchmark the performance of GMP-SDBP and SBS-SDBP. The difference between the maximum AIR for SBS-SDBP or GMP-SDBP and the maximum AIR for DBP-CG (resp. DBP-iidG) is shown in dashed (resp. solid) lines. The gains of GMP-SDBP (resp. SBS-SDBP) are shown using red circle (resp. black diamond) markers. All the curves have a maximum at some intermediate point as the gain between SDBP and DBP will be zero at small N (both AIRs saturate to the maximum value), and at large N , both AIRs vanish. The gains are lower for both GMP-SDBP and SBS-SDBP when DBP-CG is used as a benchmark technique. This is expected, and is in line with the conclusions of [16], as DBP-CG has better AIR than DBP-iidG. We can observe that using GMP-SDBP, 0.7 bits/symbol higher AIR is obtained compared to DBP-CG.

The span length is varied by keeping $N = 30$ for the 64-QAM FBG link. Similarly to Fig. 3, the highest AIR is computed, and the gains over DBP-CG are plotted in Fig. 4. We plot the gains for 14 GBd (diamond markers) and 28 GBd (circle markers) and observe that the gains for 14 GBd are higher than for 28 GBd. This behavior was observed also in our previous research [22], [30]. The SDBP accounts for signal-noise interactions, which decreases as we increase the symbol rate. Therefore, the performance of SDBP approaches DBP with increasing symbol rate.

C. Discussion

The computation of lower bounds on the MI using either an AFC or an ABC are two different ways with their own advantages and disadvantages. If an AFC is available, then $q(\mathbf{y})$ has to be calculated first, which may involve some integrals, and (4) is used to lower-bound the MI. Lower bounds on the MI can be obtained using an ABC by using any conditional distribution $r(\mathbf{x}|\mathbf{y})$, i.e., by removing the constraint that $r_q(\mathbf{x}|\mathbf{y})$ in (8) is induced by an AFC. That is, if an ABC is available, then (10) can be used to lower-bound the MI without explicitly finding an AFC. The true MI can in theory be obtained either by maximizing over all possible AFCs in (4), or by maximizing over all possible ABCs in (10). In this paper, we obtained ABCs from the SDBP algorithm and results indicate that GMP-SDBP gives the tightest lower bound on the MI compared to DBP-iidG, DBP-CG, and SBS-SDBP.

We will discuss the extension of the results for dual polarization and comment on the complexity. Firstly, computation of AIRs through the ABC using (21) is applicable for dual polarization also. SBS-SDBP has been developed for dual polarization [22] and GMP-SDBP can be extended to dual polarization. A single polarization was used for computing AIRs with GMP-SDBP in this paper for computational simplicity. We note that the polarization mode dispersion for a dual-polarization transmitter degrades the performance of both DBP and SDBP, as observed in [22], and hence the AIRs of Fig. 2 will be lowered. However, we conjecture that the relative

gains of SDBP compared to DBP would be similar to what we have shown in Fig. 2 for a single polarization. Secondly, the number of segments per span used in the simulation of the fiber using SSFM is the same as the ones used for DBP and SDBP. There exist many low-complexity variations of DBP, where the number of segments is optimized for real-time implementation. Low-complexity variations of SDBP can be derived, e.g., by optimizing the number of particles or segments per span [22].

Tighter lower bounds may possibly be obtained than those reported in the paper. Here we present two different methods. Firstly, in the GMP-SDBP, we used a linear Gaussian message passing algorithm to account for the effect of the pulse shaper, which may not be an optimal strategy for the FOC. We conjecture that when the distribution is represented in a particle form, techniques other than linear Gaussian message passing might yield even tighter bounds than those presented in the paper. Secondly, tighter bounds on the MI can be obtained by extending the principle used from AFC to ABC to a more general technique. The required property to lower-bound the MI using an AFC and also an ABC is $D \geq 0$ in (8) and (12). This principle can be extended by allowing $p(\mathbf{y})$ in (8) to be an arbitrary probability density function over \mathbf{y} that is not necessarily induced by $p(\mathbf{y}|\mathbf{x})$ or $q(\mathbf{y}|\mathbf{x})$ [21].

VI. CONCLUSION

Traditionally, lower bounds on the MI were computed using an AFC. In this paper, we computed lower bounds using an ABC for the first time for the FOC. These bounds are achievable by a maximum a posteriori detector based on the ABC. Two different distributions obtained through the SDBP algorithm are used as ABCs for estimation of AIRs. Both these distributions have better AIR in comparison to the state-of-the-art method of using DBP. Through simulations, it was also found that up to 0.7 bit/symbol higher AIR is obtained using GMP-SDBP compared to DBP. This means that in comparison to the DBP approach, tighter lower bounds can be obtained using the SDBP approach.

VII. ACKNOWLEDGMENTS

The authors would like to thank Rahul Devassy, Kamran Keykhosravi, Assoc. Prof. Giulio Colavolpe, Tobias Fehenberger, and Dr. Amina Piemontese for fruitful discussions.

REFERENCES

- [1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3/4, pp. 379–423/623–656, 1948.
- [2] P. P. Mitra and J. B. Stark, "Nonlinear limits to the information capacity of optical fibre communications," *Nature*, vol. 411, no. 6841, pp. 1027–1030, 2001.
- [3] R.-J. Essiambre, G. Kramer, P. J. Winzer, G. J. Foschini, and B. Goebel, "Capacity limits of optical fiber networks," *J. Lightw. Technol.*, vol. 28, no. 4, pp. 662–701, 2010.
- [4] A. Ellis, J. Zhao, and D. Cotter, "Approaching the non-linear Shannon limit," *J. Lightw. Technol.*, vol. 28, no. 4, pp. 423–433, 2010.
- [5] E. Agrell, A. Alvarado, G. Durisi, and M. Karlsson, "Capacity of a nonlinear optical channel with finite memory," *J. Lightw. Technol.*, vol. 32, no. 16, pp. 2862–2876, 2014.
- [6] G. Kramer, M. I. Yousefi, and F. R. Kschischang, "Upper bound on the capacity of a cascade of nonlinear and noisy channels," in *Information Theory Workshop (ITW)*, 2015.
- [7] E. Agrell, G. Durisi, and P. Johannisson, "Information-theory-friendly models for fiber-optic channels: A primer," in *Information Theory Workshop (ITW)*, 2015.
- [8] L. Wan, S. Tsai, and M. Almgren, "A fading-insensitive performance metric for a unified link quality model," in *Wireless Communications and Networking Conference (WCNC)*, 2006.
- [9] M. Franceschini, G. Ferrari, and R. Raheli, "Does the performance of LDPC codes depend on the channel?" *IEEE Trans. Commun.*, vol. 54, no. 12, pp. 2129–2132, 2006.
- [10] K. Brueninghaus, D. Astely, T. Salzer, S. Visuri, A. Alexiou, S. Karger, and G.-A. Seraji, "Link performance models for system level simulations of broadband radio access systems," in *International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2005.
- [11] A. Leven, F. Vacondio, L. Schmalen, S. T. Brink, and W. Idler, "Estimation of soft FEC performance in optical transmission experiments," *IEEE Photon. Technol. Lett.*, vol. 23, no. 20, pp. 1547–1549, 2011.
- [12] I. B. Djordjevic, B. Vasic, M. Ivkovic, and I. Gabitov, "Achievable information rates for high-speed long-haul optical transmission," *J. Lightw. Technol.*, vol. 23, no. 11, pp. 3755–3763, 2005.
- [13] G. Colavolpe, T. Foggi, A. Modenini, and A. Piemontese, "Faster-than-Nyquist and beyond: how to improve spectral efficiency by accepting interference," *Optics Express*, vol. 19, no. 27, pp. 26600–26609, 2011.
- [14] M. Secondini, E. Forestieri, and G. Prati, "Achievable information rate in nonlinear WDM fiber-optic systems with arbitrary modulation formats and dispersion maps," *J. Lightw. Technol.*, vol. 31, no. 23, pp. 3839–3852, 2013.
- [15] T. Fehenberger, A. Alvarado, P. Bayvel, and N. Hanik, "On achievable rates for long-haul fiber-optic communications," *Optics Express*, vol. 23, no. 7, pp. 9183–9191, 2015.
- [16] T. A. Eriksson, T. Fehenberger, P. A. Andrekson, M. Karlsson, N. Hanik, and E. Agrell, "Impact of 4D channel distribution on the achievable rates in coherent optical communication experiments," *J. Lightw. Technol.*, vol. 34, no. 9, pp. 2256–2266, 2016.
- [17] G. Liga, A. Alvarado, E. Agrell, and P. Bayvel, "Information rates of next-generation long-haul optical fiber systems using coded modulation," *arXiv:1606.01689 [cs.IT]*, 2016. [Online]. Available: <https://arxiv.org/abs/1606.01689>
- [18] A. Ganti, A. Lapidoth, and I. E. Telatar, "Mismatched decoding revisited: general alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, 2000.
- [19] D. M. Arnold, H.-A. Loeliger, P. O. Vontobel, A. Kavčić, and W. Zeng, "Simulation-based computation of information rates for channels with memory," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3498–3508, 2006.
- [20] N. Merhav, "Universal decoding for memoryless Gaussian channels with a deterministic interference," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1261–1269, 1993.
- [21] P. Sadeghi, P. O. Vontobel, and R. Shams, "Optimization of information rate upper and lower bounds for channels with memory," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 663–688, 2009.
- [22] N. V. Irukulapati, H. Wymeersch, P. Johannisson, and E. Agrell, "Stochastic digital backpropagation," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 3956–3968, 2014.
- [23] H. Wymeersch, N. V. Irukulapati, I. A. Sackey, P. Johannisson, and E. Agrell, "Backward particle message passing," in *Proc. Signal Processing Advances in Wireless Communications (SPAWC)*, 2015.
- [24] A. Radošević, D. Fertoni, T. M. Duman, J. G. Proakis, and M. Stojanovic, "Bounds on the information rate for sparse channels with long memory and i.i.d. inputs," *IEEE Trans. Commun.*, vol. 59, no. 12, pp. 3343–3352, 2011.
- [25] T. R. M. Fischer, "Some remarks on the role of inaccuracy in Shannon's theory of information transmission," *Transactions of the Eighth Prague Conference*, pp. 211–226, 1978.
- [26] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, 1994.
- [27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, U.S.A.: Wiley, 2006.
- [28] D. Arnold and H.-A. Loeliger, "On the information rate of binary-input channels with memory," in *IEEE Int. Conf. on Communications*, 2001.
- [29] N. V. Irukulapati, D. Marsella, P. Johannisson, M. Secondini, H. Wymeersch, E. Agrell, and E. Forestieri, "On maximum likelihood sequence detectors for single-channel coherent optical communications," in *Proc. European Conference on Optical Communication (ECOC)*, 2014.
- [30] N. V. Irukulapati, D. Marsella, P. Johannisson, E. Agrell, M. Secondini, and H. Wymeersch, "Stochastic digital backpropagation with residual

- memory compensation,” *J. Lightw. Technol.*, vol. 34, no. 2, pp. 566–572, 2016.
- [31] G. Liga, A. Alvarado, E. Agrell, M. Secondini, R. I. Killey, and P. Bayvel, “Optimum detection in presence of nonlinear distortions with memory,” in *Proc. European Conference on Optical Communication (ECOC)*, 2015.
- [32] E. Agrell, A. Alvarado, and F. R. Kschischang, “Implications of information theory in optical fibre communications,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2062, 2016.
- [33] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. Ping, and F. R. Kschischang, “The factor graph approach to model-based signal processing,” *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1295–1322, 2007.
- [34] Q. Zhang and M. I. Hayee, “Symmetrized split-step Fourier scheme to control global simulation accuracy in fiber-optic communication systems,” *J. Lightw. Technol.*, vol. 26, no. 2, pp. 302–316, 2008.
- [35] F. Rusek and D. Fertonani, “Bounds on the information rate of intersymbol interference channels based on mismatched receivers,” *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1470–1482, 2012.